From Recall to Reasoning: Automated Question Generation for Deeper Math Learning through Large Language Models

Yongan Yu^D, Alexandre Krantz^D, and Nikki G. Lobczowski^D

McGill University, Quebec, Canada

{yongan.yu, alexandre.krantz}@mail.mcgill.ca, nikki.lobczowski@mcgill.ca

Abstract. Educators have started to turn to Generative AI (GenAI) to help create new course content, but little is known about how they should do so. In this project, we investigated the first steps for optimizing content creation for advanced math. In particular, we looked at the ability of GenAI to produce high-quality practice problems that are relevant to the course content. We conducted two studies to: (1) explore the capabilities of current versions of publicly available GenAI and (2) develop an improved framework to address the limitations we found. Our results showed that GenAI can create math problems at various levels of quality with minimal support, but that providing examples and relevant content results in better quality outputs. This research can help educators decide the ideal way to adopt GenAI in their workflows, to create more effective educational experiences for students.

Keywords: Large Language Models \cdot Educational Question Generation \cdot Bloom's Taxonomy \cdot Webb's Depth of Knowledge \cdot Retrieval Augmented Generation

1 Introduction

With the rapid advancements in LLMs, AI has significantly impacted educational development [26,2], transforming how academic content is created and delivered. Recent studies demonstrate LLMs can enhance learning experiences across various scenarios [6,17]. Particularly, recent studies have explored AI applications, including personalized suggestions [32], and learning behavior analysis [3].

While over 60% of educators have experimented with ChatGPT, less than 20% feel adequately prepared to integrate it effectively [13]. One application of LLMs that could significantly support teachers is Question Generation (QG). A high-quality LLM QG tool could significantly reduce the workload of educators [4], as it would free up the time spent on creating problem sets and answers. It could also result in more practice problems for students, enhancing their learning experience. From a technical standpoint, QG is an existing sub-field of natural language processing (NLP), focused on enabling the automated creation of educational content directly from reference material, such as textbooks [7].

2 Y. Yu et al.

Despite the transformative potential of AI in education, its integration remains underutilized, especially in automated question generation. A need-finding study by Wang et al. (2023) [25] reveals that educators often express reservations about adopting AI tools, citing concerns about the relevance and quality of AI-generated content. Moreover, existing automatic QG tools are not widely used in classrooms due to their limited range in types and difficulty levels [15]. Most systems primarily produce simple recall questions, failing to sufficiently challenge students or promote deeper cognitive processing.

Problem Statement Our study investigates the gap between AI capabilities and effective educational implementation by answering: **RQ1**: What impact does increased contextual information have on the quality and cognitive depth of LLM-generated questions? and **RQ2**: How can LLM-based question generation systems be designed to produce questions across varying levels of cognitive depth? We hypothesize that developing a context-aware AI framework integrating established educational taxonomies [14] will enable the generation of high-quality, diverse, and cognitively appropriate questions that align closely with specific educational objectives and content. Our goal is to bridge the gap between AI capabilities and educational needs, potentially increasing educator confidence and the adoption of AI tools in classrooms for improved educational outcomes.

2 Related-Work

The rapid rise of GenAI, exemplified by ChatGPT's release in late 2022, has significantly impacted the educational landscape [19], with the potential to automate 20-40% of teachers' administrative tasks [6]. However, current GenAI models show limitations in math education. For instance, while effective for retrieving theorems, they struggle as conversational tutors and make errors even with elementary problems [9]. These issues stem from a lack of logical reasoning [20], highlighting the need for specialized AI systems in math education.

QG systems have evolved from simple recall questions [28] to more sophisticated models like "QG-Net" [29], which uses recurrent neural networks to generate quiz questions from educational content. However, math-specific QG faces unique challenges, including limited benchmark datasets [31] that can lead to overfitting and question redundancy [12]. Our study aims to address these limitations by creating higher-quality math questions while building AI systems that can truly adapt to students' learning journeys.

3 Study 1: Exploring Current Capabilities of GenAI

We explored how GenAI could enhance educational practices, particularly in intelligent tutoring systems and automated quizzing [24]. Our goal was to integrate GenAI seamlessly into student and teacher workflows, ensuring reliability and context awareness. To achieve this, we investigated whether GenAI could generate comprehension questions relevant to course content and analyzed their cognitive depth using Bloom's Taxonomy [1].

Table 1. Performance comparisons for three context scenarios. Note: Values are presented as mean \pm standard deviation.

	Context Scenario	Relevance	Depth	Correctness
-	Minimal	1.00 ± 0.00	2.60 ± 1.14	0.60 ± 0.55
	Moderate	1.00 ± 0.00	2.60 ± 1.82	0.80 ± 0.45
	Comprehensive	0.80 ± 0.45	2.40 ± 2.07	1.00 ± 0.00

Study Design We tested three "level of context" scenarios with progressively more information provided to the GenAI. Using a mathematical logic course covering topics such as satisfiability and the compactness theorem, we designed prompts instructing the model to generate five comprehension questions with answers. We employed GEMINI-1.5-PRO [22] with a temperature of 0 [33] to minimize output variations. In Scenario 1 (minimal), the instructor provided only the course syllabus and a brief topic summary; in Scenario 2 (moderate), they added their notes for a specific class session; and in Scenario 3 (comprehensive), they included the syllabus, class notes, and references covering course material.

Evaluation Metrics A math student with expertise in logic evaluated outputs via three metrics: relevance (binary score indicating whether the question was within lesson scope), depth (score from 0-6 corresponding to Bloom's Taxonomy levels), and correctness (binary score assessing answer accuracy).

Findings from Initial GenAI Testing In this preliminary experimentation, our analysis revealed that GenAI indeed can create content that is relevant and high-quality with input support. Results from testing were surprising in the level of relevance because there was a declining trend in the relevance level of the output as we added additional context. However, we found the depth of the questions remained relatively constant, and the correctness of the answers generated improved as more context was provided, reaching perfect accuracy with comprehensive context.

Advancing to the Next Stage: From Bloom's Taxonomy to DOK Our initial investigation revealed that additional context improved question correctness but not relevance, as the GenAI sometimes extrapolated and went beyond the bounds of what was taught in the course (i.e., creating "hallucinations"). Since instructors can generally verify answer correctness, we prioritized improving relevance in our next phase. We considered approaches like retrievalaugmented generation (RAG) [16] to better ground outputs in provided materials, which could potentially resolve this issue by forcing the AI to source its generation from a "chunk" of the materials provided. Additionally, while Bloom provided a starting framework, we found Webb's depth of knowledge (DOK) [30] framework to be better suited for math education. DOK emphasizes task complexity and contextual knowledge application rather than just cognitive processes, aligning more effectively with mathematical problem-solving and curriculum standards. This framework enables a more precise mapping of question difficulty to the cognitive processes involved in mathematical reasoning, from basic recall to complex problem-solving.

4 Y. Yu et al.



Fig. 1. Overview of the proposed framework with two core components

4 Study 2: Developing an Improved Framework

Building upon prior findings, we introduced QG-DOK, a question generation framework integrating RAG with Webb's Depth of Knowledge to generate contextaware questions with varying cognitive depth levels, addressing limitations identified in our initial phase. Thus, our second research objective explored how GenAI can generate questions of varying difficulty, making them more adaptable to teachers in math education. Our system, illustrated in Figure 1, comprises two core components.

RAG Framework QG-RAG was implemented using a naive RAG framework [10] and integrated with DOK:

- Data preprocessing: We gathered and refined mathematical content from textbooks, tutorials, and practice problems, then transformed it into vector representations through an embedding model. These vectors were stored in a database to facilitate efficient retrieval of semantically relevant materials.
- Augmented generation: When a user submits a query, the system retrieves relevant content from the vector database to provide contextual grounding. To adjust difficulty and cognitive depth, we incorporated four DOK levels into our question generation process, as shown in Figure 2 (A): Recall and Reproduction (level-1): retrieving basic facts, definitions, and formulas with minimal cognitive effort; Skills and Concepts (level-2) selecting appropriate methods and organizing information to solve routine problems; Strategic Thinking (level-3): reasoning, planning, and applying concepts in nonroutine scenarios; Extended Thinking (level-4): making connections across concepts and solving complex, multi-step problems.

Question Generation for Deeper Math Learning via LLMs

Define DOK A	RAG+Response B
DOK_level_1 = { {description : {DOK_level_1_description}} {example : {example_context}} {math-example : {math_question_context}} {reasoning : {reasoning_context}} }	Enter Your Query: {query_text} Choose DOK level: {1-4} Reference = {Retrieved_context} Passage = The question is

Fig. 2. (A) Level-1 prompt template example (B) User input interface

 Table 2. Performance comparison of LLMs across three evaluation metrics in question

 generation. Bold and <u>underline</u> indicate the highest and second-highest scores

		R	elevance	DOK	alignment	Appr	opriateness	PINC
_		DOK	DOK+RAG	DOK	DOK+RAG	DOK	DOK+RAG	I
GPT-40	Level 1 Level 2 Level 3 Level 4	0.85 0.81 0.79 0.78	0.79 0.85 0.81 0.80	$\begin{array}{c} \underline{0.81} \\ 0.30 \\ 0.52 \\ 0.29 \end{array}$	0.78 0.73 0.71 0.61	$\begin{array}{c} 0.80 \\ \hline 0.72 \\ 0.74 \\ 0.70 \end{array}$	0.91 0.90 0.95 0.82	0.94 0.93 0.92 0.90
	Average	0.81	0.81	0.48	0.71	0.74	0.90	0.92
EPSEEK-V3	Level 1 Level 2 Level 3 Level 4	$ \begin{array}{c} 0.84 \\ 0.80 \\ 0.77 \\ 0.66 \end{array} $	$ \begin{array}{r} 0.82 \\ \underline{0.84} \\ 0.80 \\ 0.71 \end{array} $	0.80 0.75 0.63 0.39	$0.68 \\ 0.72 \\ 0.78 \\ 0.59$	0.82 0.74 0.72 0.71	$ \begin{array}{r} 0.89 \\ 0.88 \\ 0.85 \\ 0.80 \end{array} $	0.94 0.92 0.90 0.92
Ē	Average	0.77	0.79	0.64	0.69	0.75	0.86	0.92
GEMINI-1.5	Level 1 Level 2 Level 3 Level 4	0.80 0.75 0.72 0.63	0.75 0.78 0.74 0.73	0.75 0.49 0.52 0.46	0.77 0.68 0.75 0.58	0.76 0.68 0.69 0.65	0.85 0.82 0.88 0.75	0.92 0.90 0.89 0.92

Implementation Details As shown in Figure 1, we embedded a corpus of mathematical content using the TEXT-BEDDING-ADA-002 model. To enhance semantic relevance, documents were segmented into fixed-size chunks using a sliding-window approach [27]. UI functionality is shown in Figure 2 (B), users were prompted to input two key pieces of information: the mathematical concept they wish to explore and the desired DOK level. In second study, we evaluated three off-the-shelf LLMs for question generation with default temperature settings, including GPT-40 [21], DEEPSEEK-V3 [5] and GEMINI-1.5-PRO [11].

Evaluation Metrics To assess our QG-DOK framework, we compared two implementations: (1) DOK, providing only DOK level definitions, and (2) DOK+ RAG, which retrieved relevant examples from a vector database. We evaluated using G-Eval [18] to measure relevance, DOK alignment, and appropriateness. Additionally, we incorporated the Paraphrase n-gram Change (PINC) score [23] to quantify lexical diversity.

Findings from Improved Work Our evaluation demonstrated that DOK+ RAG consistently outperformed DOK-only across all tested LLMs, particularly for higher-order thinking skills (DOK Levels 3 & 4). DOK+RAG improved both relevance and appropriateness scores, with GPT-40 showing the most significant gains in appropriateness. Although DOK alignment showed mixed results, DOK+RAG generally improved depth accuracy at higher cognitive levels. Also,

5

6 Y. Yu et al.

high PINC scores (average 0.92) indicated strong lexical diversity in question rephrasing. Despite these improvements, challenges remained. The depth alignment at Level 2 was inconsistent across the models, suggesting that LLMs struggle with categorizing mid-level cognitive complexity. We also identified persistent issues in mathematical notation handling that would benefit from LaTeX-based representation in future implementations.

5 Conclusion and Discussion

Through two interconnected studies, our research provides insights into GenAI's potential in math education. Study 1 revealed that while GenAI can produce relevant questions using Bloom's Taxonomy, it struggles with higher cognitive levels and tends to generate plausible yet incorrect information when given more context. Building on these findings, Study 2 introduced the QG-DOK framework, integrating Webb's DOK levels with RAG. By leveraging resources that educators are already familiar with, both quality and depth of generated questions improved. Our findings support earlier research suggesting that AI can effectively generate educational content but requires careful design to ensure cognitive depth and relevance [6]. The improvements in Study 2 address limitations identified in Study 1, particularly in generating deeper thinking questions.

6 Limitations

Although our results highlight the potential of GenAI in educational content creation, a few limitations remain. Bloom's taxonomy, though often used to categorize cognitive understanding, has been critiqued for oversimplifying the interconnected nature of learning [8], reflecting concerns in our study. Switching to DOK to inform the AI for question generation helps alleviate this to a degree to meet our exploratory goals, but it is still possible that simply describing each level is not enough instruction for the AI to create an appropriate question. Future work could explore how educators design problems to derive a step-bystep framework for AI-driven question generation.

We also acknowledge the constraints of our data input and evaluation methods. We used a single reference content as context for the AI in each study, and a single human evaluator. Given the exploratory nature of this study, though, it was important to limit the parameters to best explain variations in the quality. Moving forward, researchers can test and evaluate the output more systematically, given that our findings have highlighted the capabilities (and limitations) of current genAI. As such, our work serves as an important preliminary step in advancing question generation through AI for advanced math.

Acknowledgements This work was supported by an exploratory Interdisciplinary Research Development award from McCAIS at McGill University.

References

- Anderson, L.W., Krathwohl, D.R.: A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc. (2001)
- Antu, S.A., Chen, H., Richards, C.K.: Using llm (large language model) to improve efficiency in literature review for undergraduate research. LLM@ AIED pp. 8–16 (2023)
- Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y., Hussain, A.: Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies 28(1), 905–971 (2023)
- De Kuthy, K., Kannan, M., Ponnusamy, H.S., Meurers, D.: Towards automatically generating questions under discussion to link information and discourse structure. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5786–5798 (2020)
- DeepSeek: DeepSeek-v3 technical report (2024), https://arxiv.org/abs/2412. 19437
- Denny, P., Prather, J., Becker, B.A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B.N., Santos, E.A., Sarsa, S.: Computing education in the era of generative ai. Communications of the ACM 67(2), 56–67 (2024)
- Elkins, S., Kochmar, E., Cheung, J.C., Serban, I.: How teachers can use large language models and bloom's taxonomy to create educational quizzes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 23084–23091 (2024)
- 8. Fadul, J.A.: Collective learning: Applying distributed cognition for collective intelligence. International Journal of Learning **16**(4) (2009)
- Frieder, S., Pinchetti, L., Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P., Berner, J.: Mathematical capabilities of chatgpt. Advances in neural information processing systems 36 (2024)
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
- Gemini: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), https://arxiv.org/abs/2403.05530
- 12. Guo, S., Liao, L., Li, C., Chua, T.S.: A survey on neural question generation: Methods, applications, and prospects. arXiv preprint arXiv:2402.18267 (2024)
- 13. Hojeij, Z., Kuhail, M.A., ElSayary, A.: Investigating in-service teachers' views on chatgpt integration. Interactive Technology and Smart Education (2024)
- 14. Irvine, J.: Taxonomies in education: Overview, comparison, and future directions. Journal of Education and Development 5(2), 1 (2021)
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education 30, 121–204 (2020)
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33, 9459–9474 (2020)
- Li, H., Xu, T., Zhang, C., Chen, E., Liang, J., Fan, X., Li, H., Tang, J., Wen, Q.: Bringing generative ai to adaptive learning in education. arXiv preprint arXiv:2402.14601 (2024)

- 8 Y. Yu et al.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634 (2023)
- 19. Lo, C.K.: What is the impact of chatgpt on education? a rapid review of the literature. Education Sciences **13**(4), 410 (2023)
- McCarthy, J.: Artificial intelligence, logic, and formalising common sense. Machine Learning and the City: Applications in Architecture and Urban Design pp. 69–90 (2022)
- 21. OpenAI: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
- Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
- Scaria, N., Dharani Chenna, S., Subramani, D.: Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In: International Conference on Artificial Intelligence in Education. pp. 165–179. Springer (2024)
- Shute, V.J., Psotka, J.: 19. intelligent tutoring systems: Past, present, and future (1994)
- 25. Wang, D., Tao, Y., Chen, G.: Artificial intelligence in classroom discourse: A systematic review of the past decade. International Journal of Educational Research 123, 102275 (2024). https://doi.org/https://doi.org/10.1016/ j.ijer.2023.102275, https://www.sciencedirect.com/science/article/pii/ S0883035523001386
- Wang, N., Johnson, M.: Ai education for k-12: Connecting ai concepts to high school math curriculum. Workshop on Education in Artificial Intelligence K-12, 28th International Joint Conference on Artificial Intelligence https://par.nsf. gov/biblio/10440200
- 27. Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., et al.: Searching for best practices in retrieval-augmented generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 17716–17736 (2024)
- Wang, X., Fan, S., Houghton, J., Wang, L.: Towards process-oriented, modular, and versatile question generation that meets educational needs. arXiv preprint arXiv:2205.00355 (2022)
- Wang, Z., Lan, A.S., Nie, W., Waters, A.E., Grimaldi, P.J., Baraniuk, R.G.: Qg-net: a data-driven question generation model for educational content. In: Proceedings of the fifth annual ACM conference on learning at scale. pp. 1–10 (2018)
- Webb, N.L.: Depth-of-knowledge levels for four content areas. Language Arts 28(March), 1–9 (2002)
- Wu, H., Hui, W., Chen, Y., Wu, W., Tu, K., Zhou, Y.: Conic10k: a challenging math problem understanding and reasoning dataset. arXiv preprint arXiv:2311.05113 (2023)
- 32. Xiong, Z., Li, H., Liu, Z., Chen, Z., Zhou, H., Rong, W., Ouyang, Y.: A review of data mining in personalized education: Current trends and future prospects. Frontiers of Digital Education 1(1), 26–50 (2024)
- 33. Zhu, Y., Li, J., Li, G., Zhao, Y., Jin, Z., Mei, H.: Hot or cold? adaptive temperature sampling for code generation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 437–445 (2024)